

Porting Erasmus Computing Grid, (Condor enabled) applications for EDGeS

**Erasmus Medical Centre
Hogeschool Rotterdam - CMI**

Tobias A. Knoch

Head Biophysical Genomics,
Head
(Erasmus Computing) Grid Office

Erasmus Medical Centre:
Dr. Molewaterplein 50,
3015 GE Rotterdam

E-mail: TA.Knoch@taknoch.org

Luc V. de Zeeuw

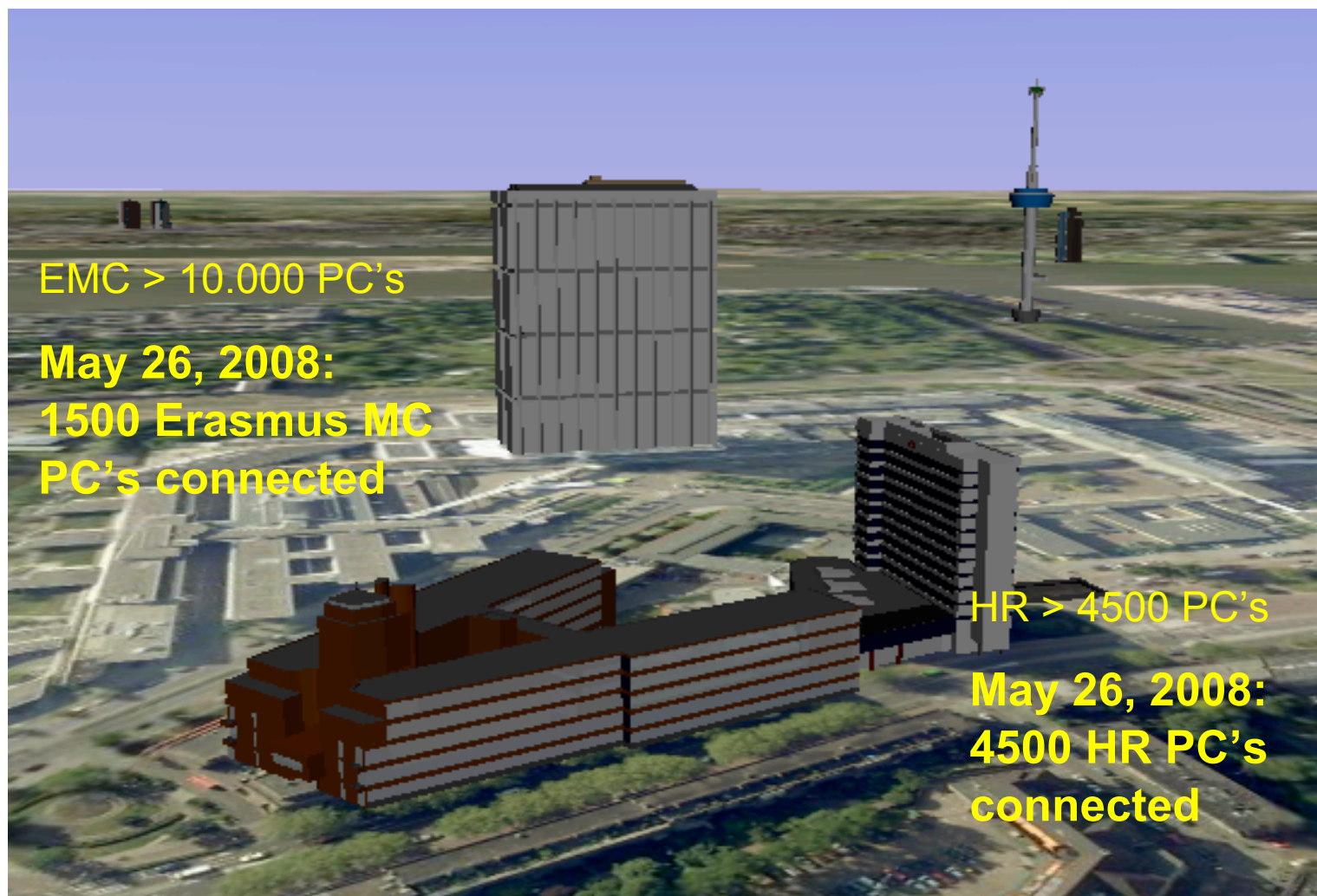
Lecturer
Coordinator Telematics department
Head
(Erasmus Computing) Grid Office

Hogeschool Rotterdam /
Communication Media and Information
Technologie (CMI)
G.J. de Jonghweg 4-6,
3015 GG Rotterdam
E-mail: L.V.de.Zeeuw@hro.nl

Contents

- Erasmus Computing Grid
- Applications for EDGeS

Erasmus Computing Grid



Bron: <http://rotterdamlandmarks.hoppinger.com/>



Condor

High Throughput Computing

decimal.hro.nl - PuTTY

HR-00304883EE	WINNT51	INTEL	Owner	Idle	0.000	510[?????]	
HR-00304883F3	WINNT51	INTEL	Owner	Idle	0.050	254	0+00:07:50
HR-00304884A1	WINNT51	INTEL	Owner	Idle	0.010	1022	0+00:42:50
HR-0030488639	WINNT51	INTEL	Owner	Idle	0.020	1022	0+00:37:51
HR-0030488753	WINNT51	INTEL	Owner	Idle	0.020	1022	0+00:42:51
HR-0030488758	WINNT51	INTEL	Owner	Idle	0.010	1022[?????]	
HR-003048875A	WINNT51	INTEL	Owner	Idle	0.010	1022[?????]	
HR-003048875A	WINNT51	INTEL	Unclaimed	Idle	0.020	510[?????]	
HR-003048875A	WINNT51	INTEL	Owner	Idle	0.000	1022[?????]	
HR-003048883C	WINNT51	INTEL	Owner	Idle	0.000	502	0+01:42:48
HR-00304888E2	WINNT51	INTEL	Owner	Idle	0.010	1022	0+00:02:51
HR-0050568A1F	WINNT51	INTEL	Owner	Idle	0.460	511	1+14:27:39
HR-0050568A2F	WINNT51	INTEL	Unclaimed	Idle	0.010	511	0+00:27:39
HR-00D0B7E317	WINNT51	INTEL	Owner	Idle	0.020	254	0+02:27:41
ORCHIDEE	WINNT51	INTEL	Owner	Idle	0.440	510	18+21:12:17
TOSHIBA	WINNT51	INTEL	Owner	Idle	0.090	503	0+02:42:53

	Total	Owner	Claimed	Unclaimed	Matched	Preempting	Backfill
INTEL/LINUX	2	2	0	0	0	0	0
INTEL/WINNT51	2653	1995	389	269	0	0	0
Total	2655	1997	389	269	0	0	0

[decimal] </home/grid/gridman>

Objectives

Erasmus Computing Grid

- Using rest and over capacity for socially relevant scientific research.

Teaching:

- Training technicians: building, management and usage of GRID infrastructures.
- Opportunity for GRID related internships.
- Students will be able to use a realistic GRID infrastructure
- Multidisciplinary GRID related projects.

Mission:

- To make GRIDs well known to as the technology to make better use of computer resources.
- To be an example for other institutions to donate their unused resources for the benefit of science

EDGeS Applications

Genetics/cell biology:

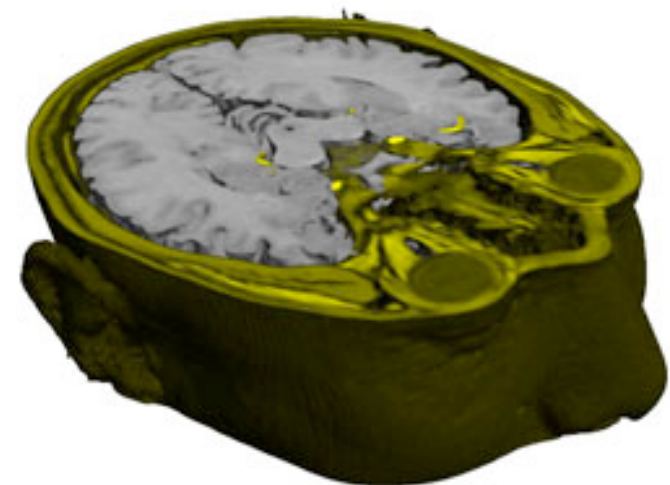
- High-Throughput Sequencing: pipeline analyses
- Sequence comparison
- Simulations of large macromolecular polymer structures

Brain research:

- MRI scan analyses

MRI-Morph

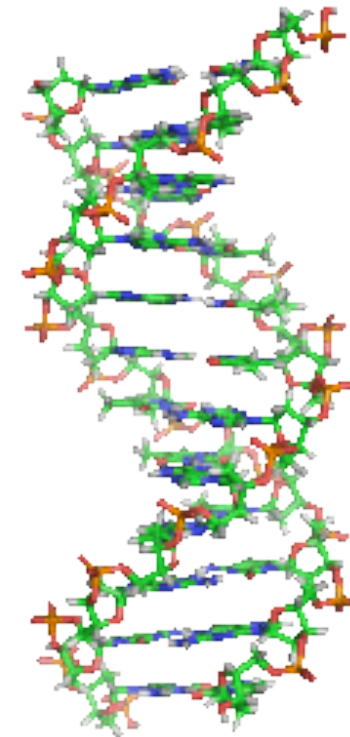
MRI-Morph: To understand the differences in brains the Rotterdam Study comprised of ~15,000 participants is taking every second year 3D-MRI scans from these participants. These are analysed and compared by MRI-Morph. From the results correlations are made to other information provided by the Rotterdam Study. The impact of this analysis is of major importance for basic research to understand brain function in general and beyond in diagnostic and treatment in relation to disease. Since the Rotterdam Study is one of the largest studies of this kind major results are expected. The ethical issues involved in this study were already taken care of by the study organizers and pose no risk for griddification



Genetics (HTS-Analysis)

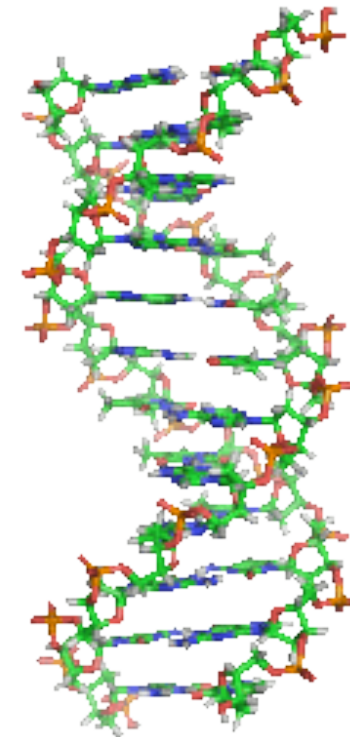
HTS-Analysis: High-Throughput Sequencing is a pipeline analysing the data coming from high-throughput sequencers. The 5-10 Terabytes originating from the sequencer comes split in relatively tiny pieces and has to be formatted, processed and mapped to a reference sequence. This is done in mainly 3 steps consisting of i) image analysis, ii) individual "shot-gun" sequence determination, and iii) assembly of the entire genomic sequence. The functioning of this pipeline is proven and is used in daily work. Due to the importance of high-throughput sequencing for research and diagnostics, this is a prime application for grid with major impact. All ethical issues in terms of data privacy were already taken care of by the providers of the material and pose no risk for griddification.

Tobias A. Knoch / Luc V. de Zeeuw



Genetics (DNA-ORG)

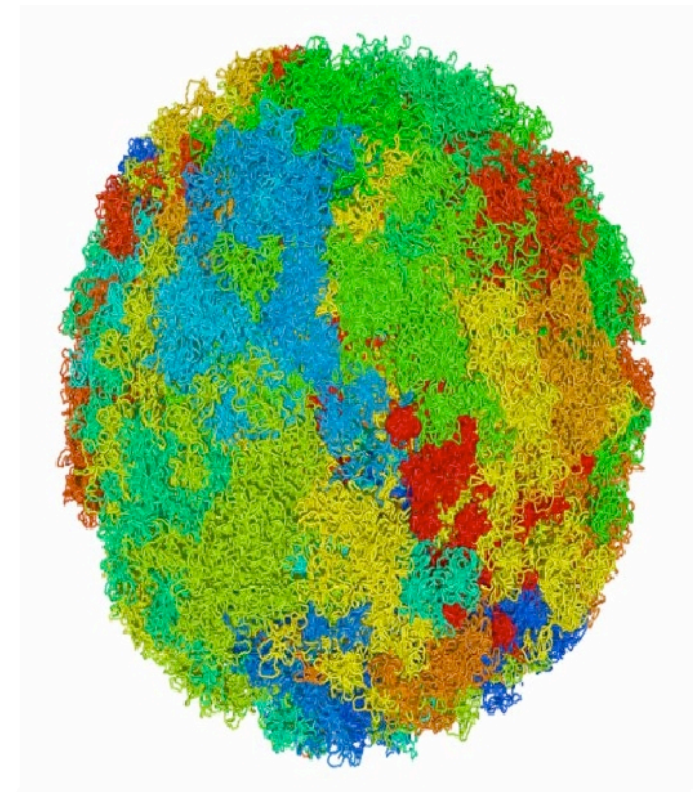
DNA-ORG: The major task in genomics is sequence comparison and pattern recognition within this sequence. DNA-ORG consists of three major and new algorithms to i) make exact comparisons between sequences, ii) find arbitrary patterns in genetic sequences, and iii) localized patterns in sequences. Currently, DNA-ORG is used to analyse these parameters in ~8,000 completely sequenced genomes with a high impact for the understanding of genomes. Beyond, with the just starting boom in HTS there are unprecedented opportunities for grid based high-performance computing. All ethical issues in terms of data privacy were already taken care of by the providers of the material and pose no risk for griddification.



VirtNuc

VirtNuc: The application VirtNuc which makes simulations of large macromolecular polymer structures. It is able to simulate extremely large structures. Currently, it is used to predict the architectural/3D organization of the human genome by simulating all chromosomes in the cell nucleus assuming the chromatin fibre as consisting of 2,400,000 million segments. With Monte-Carlo and Brownian Dynamics methods configurations of chromosomal topologies and the entire mitosis, i.e. nuclear cell-cycle are simulated and used for prediction and comparison to experiments. This is of major importance for genetic understanding, diagnostics and treatment in the framework of genetic engineering for disease treatment. No ethical issues what so ever apply.

Tobias A. Knoch / Luc V. de Zeeuw

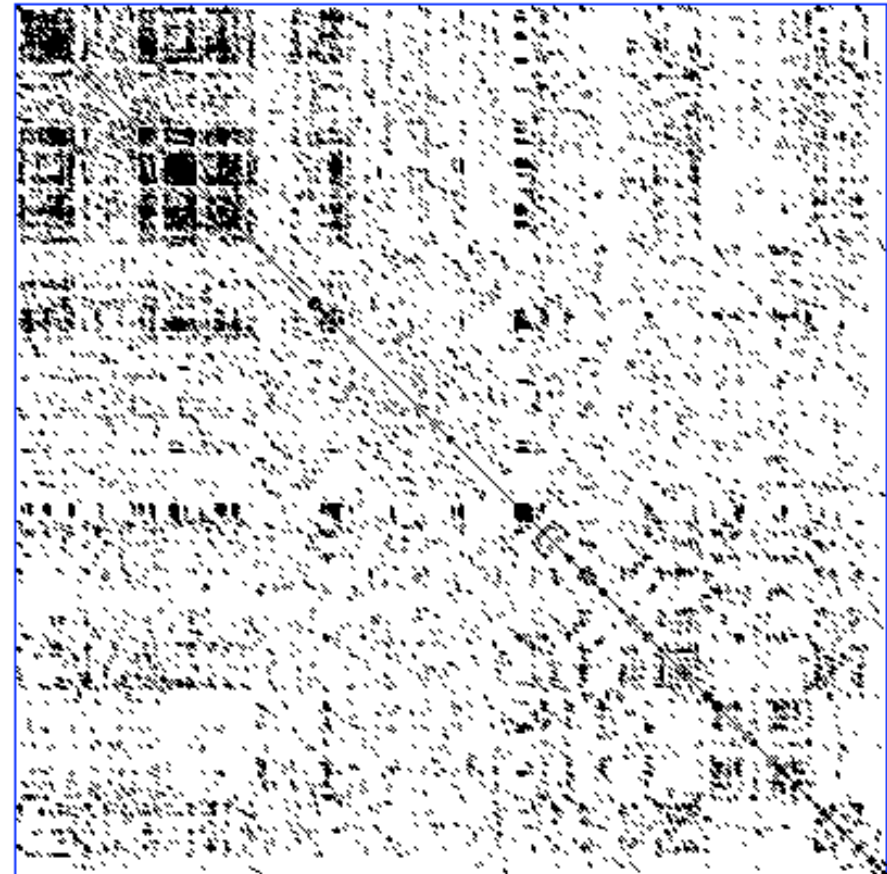
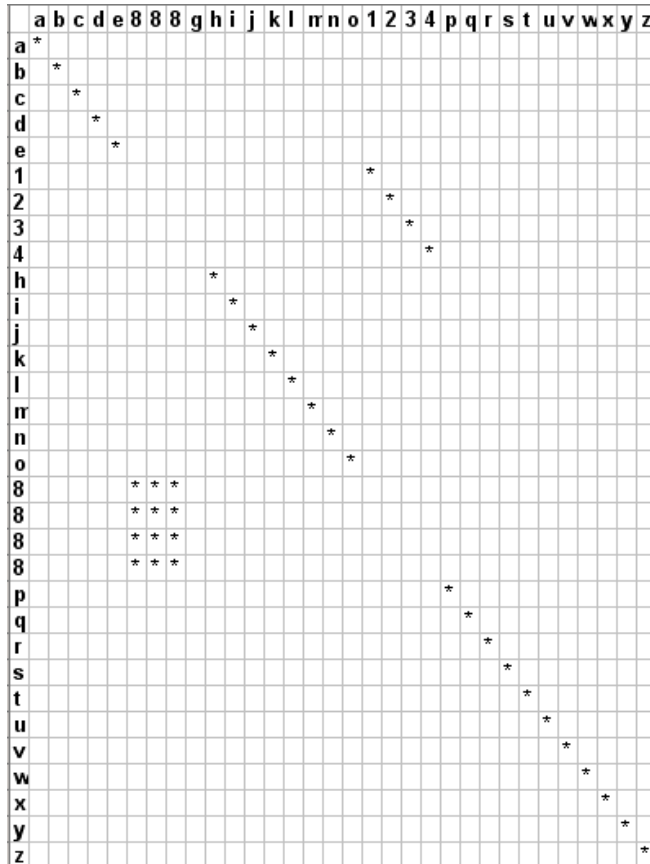


Alignment Algorithms

Time Space

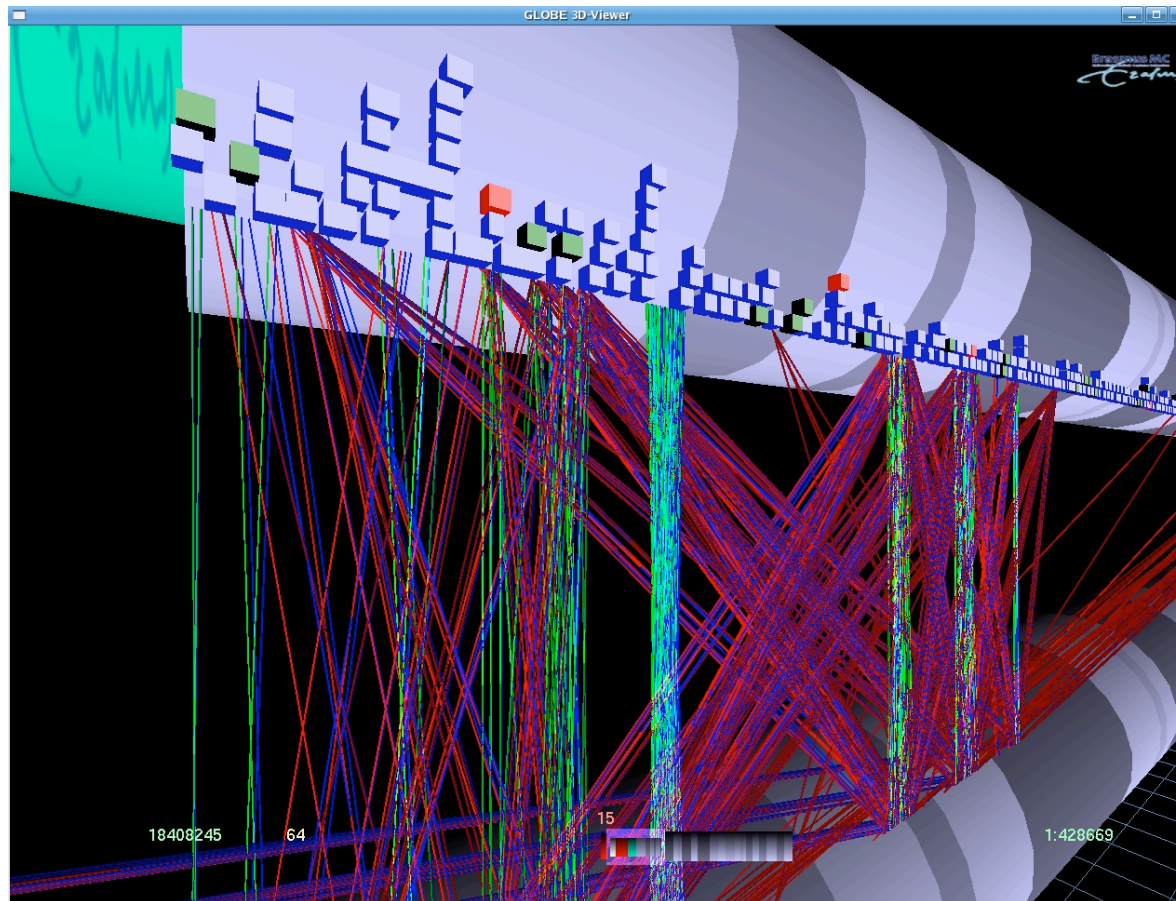
	Time	Space
Needleman-Wunsch-Gotoh	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2)$
Smith-Waterman-Gotoh	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2)$
Hirschberg-Myers-Miller	$\mathcal{O}(N^2)$	$\mathcal{O}(N)$
variant Needleman-Wunsch-Gotoh	$\mathcal{O}(N^2)$	$\mathcal{O}(\max(N, L^2))$
variant Hirschberg-Myers-Miller	$\mathcal{O}(N^2)$	$\mathcal{O}(N)$

Dot plot



Visualization of Results in DNA viewer

Prader-Willi
syndrome





Erasmus Computing Grid

—

Het Bouwen van een 20 TeraFLOP Virtuelle Supercomputer

Knoch, T. A. & de Zeeuw, L.,

Nationaal Informatica Onderwijs Congres - NIOC 2007, Amsterdam, The Netherlands, 17th - 18th April, 2007.

Abstract

Het Erasmus Medische Centrum (Erasmus MC) en Hogeschool Rotterdam (HR) zijn in 2005 een unieke samenwerking begonnen om 95% van de capaciteit op al haar computers en die van anderen beschikbaar te maken voor onderzoek en onderwijs. Deze samenwerking heeft geleid tot het Erasmus Computing GRID (ECG), een virtuele supercomputer, met op dit moment een capaciteit van 5 Tera FLOPS maar na voltooiing met een capaciteit van 20 Tera (10^{12}) FLOPS (Floting Point Operations Per Second) . Na uitvoerig testen is het ECG in productie gegaan en wordt gebruikt door tal van research groepen bij het Erasmus MC en voor onderwijs bij de HR.

Wij geven u géén diepgaande technische uiteenzettingen over grid computing, daarvoor willen we u graag verwijzen naar de uitstekende vakliteratuur en de vele bronnen op het Internet. Wél voorbeelden van toepassingen en een pleidooi om grid computing binnen het onderwijs een betere basis te geven: er is behoefte aan goed opgeleide technici die een grid infrastructuur kunnen bouwen, beheren en gebruiken.

Corresponding author email contact: TA.Knoch@taknoch.org

Keywords:

Genome, genomics, genome organization, genome architecture, structural sequencing, architectural sequencing, systems genomics, coevolution, holistic genetics, genome mechanics, genome function, genetics, gene regulation, replication, transcription, repair, homologous recombination, simultaneous co-transfection, cell division, mitosis, metaphase, interphase, cell nucleus, nuclear structure, nuclear organization, chromatin density distribution, nuclear morphology, chromosome territories, subchromosomal domains, chromatin loop aggregates, chromatin rosettes, chromatin loops, chromatin fibre, chromatin density, persistence length, spatial distance measurement, histones, H1.0, H2A, H2B, H3, H4, mH2A1.2, DNA sequence, complete sequenced genomes, molecular transport, obstructed diffusion, anomalous diffusion, percolation, long-range correlations,

fractal analysis, scaling analysis, exact yard-stick dimension, box-counting dimension, lacunarity dimension, local nuclear dimension, nuclear diffuseness, parallel super computing, grid computing, volunteer computing, Brownian Dynamics, Monte Carlo, fluorescence in situ hybridization, confocal laser scanning microscopy, fluorescence correlation spectroscopy, super resolution microscopy, spatial precision distance microscopy, auto-fluorescent proteins, CFP, GFP, YFP, DsRed, fusion protein, in vivo labelling, information browser, visual data base access, holistic viewing system, integrative data management, extreme visualization, three-dimensional virtual environment, virtual paper tool.

Literature References

- Knoch, T. A.** Dreidimensionale Organisation von Chromosomen-Domänen in Simulation und Experiment. (Three-dimensional organization of chromosome domains in simulation and experiment.) *Diploma Thesis*, Faculty for Physics and Astronomy, Ruperto-Carola University, Heidelberg, Germany, 1998, and TAK Press, Tobias A. Knoch, Mannheim, Germany, ISBN 3-00-010685-5 and ISBN 978-3-00-010685-9 (soft cover, 2nd ed.), ISBN 3-00-035857-9 and ISBN 978-3-00-0358857-0 (hard cover, 2nd ed.), ISBN 3-00-035858-7, and ISBN 978-3-00-035858-6 (DVD, 2nd ed.), 1998.
- Knoch, T. A.,** Munkel, C. & Langowski, J. Three-dimensional organization of chromosome territories and the human cell nucleus - about the structure of a self replicating nano fabrication site. *Foresight Institute - Article Archive*, Foresight Institute, Palo Alto, CA, USA, <http://www.foresight.org>, 1- 6, 1998.
- Knoch, T. A.,** Munkel, C. & Langowski, J. Three-Dimensional Organization of Chromosome Territories and the Human Interphase Nucleus. *High Performance Scientific Supercomputing*, editor Wilfried Juling, Scientific Supercomputing Center (SSC) Karlsruhe, University of Karlsruhe (TH), 27- 29, 1999.
- Knoch, T. A.,** Munkel, C. & Langowski, J. Three-dimensional organization of chromosome territories in the human interphase nucleus. *High Performance Computing in Science and Engineering 1999*, editors Krause, E. & Jäger, W., High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York, ISBN 3-540-66504-8, 229-238, 2000.
- Bestvater, F., **Knoch, T. A.,** Langowski, J. & Spiess, E. GFP-Walking: Artificial construct conversions caused by simultaneous cotransfection. *BioTechniques* 32(4), 844-854, 2002.
- Knoch, T. A. (editor),** Backes, M., Baumgärtner, V., Eysel, G., Fehrenbach, H., Göker, M., Hampl, J., Hampl, U., Hartmann, D., Hitzelberger, H., Nambena, J., Rehberg, U., Schmidt, S., Weber, A., & Weidemann, T. Humanökologische Perspektiven Wechsel - Festschrift zu Ehren des 70. Geburtstags von Prof. Dr. Kurt Egger. Human Ecology Working Group, Ruperto-Carola University of Heidelberg, Heidelberg, Germany, 2002.
- Knoch, T. A.** Approaching the three-dimensional organization of the human genome: structural-, scaling- and dynamic properties in the simulation of interphase chromosomes and cell nuclei, long- range correlations in complete genomes, *in vivo* quantification of the chromatin distribution, construct conversions in simultaneous co-transfections. *Dissertation*, Ruperto-Carola University, Heidelberg, Germany, and TAK†Press, Tobias A. Knoch, Mannheim, Germany, ISBN 3-00-009959-X and ISBN 978-3-00-009959-5 (soft cover, 3rd ed.), ISBN 3-00-009960-3 and ISBN 978-3-00-009960-1 (hard cover, 3rd ed.), ISBN 3-00-035856-9 and ISBN 978-3-00-010685-9 (DVD, 3rd ed.) 2002.
- Knoch, T. A.** Towards a holistic understanding of the human genome by determination and integration of its sequential and three-dimensional organization. *High Performance Computing in Science and Engineering 2003*, editors Krause, E., Jäger, W. & Resch, M., High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York, ISBN 3- 540-40850-9, 421-440, 2003.
- Wachsmuth, M., Weidemann, T., Müller, G., Urs W. Hoffmann-Rohrer, **Knoch, T. A.,** Waldeck, W. & Langowski, J. Analyzing intracellular binding and diffusion with continuous fluorescence photobleaching. *Biophys. J.* 84(5), 3353-3363, 2003.

- Weidemann, T., Wachsmuth, M., **Knoch, T. A.**, Müller, G., Waldeck, W. & Langowski, J. Counting nucleosomes in living cells with a combination of fluorescence correlation spectroscopy and confocal imaging. *J. Mol. Biol.* 334(2), 229-240, 2003.
- Fejes Tóth, K., **Knoch, T. A.**, Wachsmuth, M., Frank-Stöhr, M., Stöhr, M., Bacher, C. P., Müller, G. & Rippe, K. Trichostatin A induced histone acetylation causes decondensation of interphase chromatin. *J. Cell Science* 117, 4277-4287, 2004.
- Ermiler, S., Kronic, D., **Knoch, T. A.**, Moshir, S., Mai, S., Greulich-Bode, K. M. & Boukamp, P. Cell cycle-dependent 3D distribution of telomeres and telomere repeat-binding factor 2 (TRF2) in HaCaT and HaCaT-myc cells. *Europ. J. Cell Biol.* 83(11-12), 681-690, 2004.
- Kost, C., Gama de Oliveira, E., **Knoch, T. A.** & Wirth, R. Spatio-temporal permanence and plasticity of foraging trails in young and mature leaf-cutting ant colonies (*Atta spp.*). *J. Trop. Ecol.* 21(6), 677- 688, 2005.
- Winnefeld, M., Grewenig, A., Schnölzer, M., Spring, H., **Knoch, T. A.**, Gan, E. C., Rommelaere, J. & Cziepluch, C. Human SGT interacts with BAG-6/Bat-3/Scythe and cells with reduced levels of either protein display persistence of few misaligned chromosomes and mitotic arrest. *Exp. Cell Res.* 312, 2500-2514, 2006.
- Sax, U., Weisbecker, A., Falkner, J., Viezens, F., Yassene, M., Hartung, M., Bart, J., Krefting, D., **Knoch, T. A.** & Semler, S. Grid-basierte Services für die elektronische Patientenakte der Zukunft. *E- HEALTH-COM - Magazin für Gesundheitstelematik und Telemedizin*, 4(2), 61-63, 2007.
- de Zeeuw, L. V., **Knoch, T. A.**, van den Berg, J. & Grosveld, F. G. Erasmus Computing Grid - Het bouwen van een 20 TeraFLOP virtuele supercomputer. *NIOC proceedings 2007 - het perspectief van lange termijn*. editor Frederik, H. NIOC, Amsterdam, The Netherlands, 52-59, 2007.
- Rauch, J., **Knoch, T. A.**, Solovei, I., Teller, K. Stein, S., Buiting, K., Horsthemke, B., Langowski, J., Cremer, T., Hausmann, M. & Cremer, C. Lightoptical precision measurements of the Prader- Willi/Angelman Syndrome imprinting locus in human cell nuclei indicate maximum condensation changes in the few hundred nanometer range. *Differentiation* 76(1), 66-82, 2008.
- Sax, U., Weisbecker, A., Falkner, J., Viezens, F., Mohammed, Y., Hartung, M., Bart, J., Krefting, D., **Knoch, T. A.** & Semler, S. C. Auf dem Weg zur individualisierten Medizin - Grid-basierte Services für die EPA der Zukunft. *Telemedizinführer Deutschland 2008*, editor Jäckel, A. Deutsches Medizinforum, Minerva KG, Darmstadt, ISBN 3-937948-06-6, ISBN-13 9783937948065, 47-51, 2008.
- Drägestein, K. A., van Capellen, W. A., van Haren, J. Tsibidis, G. D., Akhmanova, A., **Knoch, T. A.**, Grosveld, F. G. & Galjart, N. Dynamic behavior of GFP-CLIP-170 reveals fast protein turnover on microtubule plus ends. *J. Cell Biol.* 180(4), 729-737, 2008.
- Jhunjhunwala, S., van Zelm, M. C., Peak, M. M., Cutchin, S., Riblet, R., van Dongen, J. J. M., Grosveld, F. G., **Knoch, T. A.**⁺ & Murre, C.⁺ The 3D-structure of the Immunoglobulin Heavy Chain Locus: implications for long-range genomic interactions. *Cell* 133(2), 265-279, 2008.
- Krefting, D., Bart, J., Beronov, K., Dzhimova, O., Falkner, J., Hartung, M., Hoheisel, A., **Knoch, T. A.**, Lingner, T., Mohammed, Y., Peter, K., Rahm, E., Sax, U., Sommerfeld, D., Steinke, T., Tolxdorff, T., Vossberg, M., Viezens, F. & Weisbecker, A. MediGRID - Towards a user friendly secured grid infrastructure. *Future Generation Computer Systems* 25(3), 326-336, 2008.